社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59 (40)
2005／12／22

雑音環境下で視覚情報が日本語音節認識に及ぼす効果について

星野 真人[1] 伊藤 仁[2] 木村 真弘[1] 中村 真[3]

1 株式会社 ホンダ・リサーチ・インスティチュート・ジャパン 〒351-0114 埼玉県和光市本町 8-1
2 東北大学電気通信研究所 〒980-0812 仙台市青葉区片平 2-1-1
3 株式会社 本田技術研究所和光基礎技術研究センター第 5 研究室 〒351-0114 埼玉県和光市本町 8-1

E-mail: 1 mhoshino@jp.honda-ri.com, mkimura@jp.honda-ri.com 2 itojin@riec.tohoku.ac.jp
3 Makoto_Nakamura@n.f.rd.honda.co.jp

あらまし 視覚情報による日本語音節認識能力の雑音環境下における改善効果に関して心理物理実験を通じて得られた知見を報告する.

キーワード 日本語音節認識, 雑音環境, 視覚情報, 第 1 ホルマント周波数

The Effect of Visual Information for Japanese Syllables Recognition

Masato HOSHINO[1] Masashi ITO[2] Masahiro KIMURA[1] and Makoto MAKAMURA[3]

1 Honda Research Institute Japan Co.,Ltd. 8-1 Honcho, Wako-shi, Saitama, 351-0114 Japan
2 RIEC, Tohoku University 2-1-1 Katahira, Aoba-ku, Sendai-shi, Miyagi 980-0812 Japan
3 Honda R&D Co.,Ltd. 8-1 Honcho, Wako-shi, Saitama, 351-0114 Japan

E-mail: 1 mhoshino@jp.honda-ri.com, mkimura@jp.honda-ri.com 2 itojin@riec.tohoku.ac.jp
3 Makoto_Nakamura@n.f.rd.honda.co.jp

Abstract We will report some findings regarding the effects of visual information for an improved recognition of Japanese Syllables under a noisy environment. These findings are from our phychophysical experiment.

Keyword Japanese Syllables Recognition, Noisy Environment, Visual Information, F1

1. はじめに

近年の音声認識技術は, 製品を一般ユーザーが入手できる段階まで発達している.しかしながら,実環境では雑音環境下での著しい性能劣化の問題が応用範囲を狭める要因となっている.

この問題を解決するためのアプローチに,音声だけでなく話者の顔画像も入力情報として用いる手法がある.このマルチモーダル音声認識システムは,雑音環境下での性能劣化を解決する技術として期待が大きく,いくつかの研究機関で基本技術の開発が行われている[1][2].このマルチモーダル音声認識では,画像情報からの適切な特徴抽出法と,画像と音声から得られた特徴の適切な統合方法の 2 点が大きな技術課題である.

一方,上記の技術課題を解決する上で,マルチモーダルな音声認識を行っていると考えられる人間が,視聴覚情報を用いて音声認識する際の性質について調べることは重要である.積山らは,日本語音節を発話する話者の画像だけを呈示する実験を行い,訓練をしていない通常の人間の読唇能力を調べた[3].その結果,音節単位での正答率は 15%程度だが,母音に限れば 90%近い認識率が得られた.彼らは読唇による認識誤りは主に子音にあると考え,子音をその誤りパターンから 5-6 のグループに分類した.また英語について調べた研究でも同様の結果が得られ,視覚情報から弁別できる子音グループは" phoneme" を文字って" viseme" と呼ばれている[4].

この様に人間のマルチモーダルな音声認識を調べる研究はいくつかあるが,雑音下で音声を認識する際に視覚情報がどの程度有効なのか,またその有効性のメカニズムは明確になっていない.本稿では,ロバストな音声認識のための視覚情報の有効性を定量的に評価した.具体的には,発話時の話者の動画像と雑音を付加した音声信号を刺激として,日本語音節の認識実験を行い人間の音声認識能力を調べた.この実験により得た知見をもとに,聴覚による雑音環境下における音声認識の限界,視覚情報の付加で得られる音声認識能力の改善の大きさと人間のマルチモーダル音声認識の仕組みについて考察する.

2.実験手法

2.1 呈示刺激

実験に用いた刺激は以下の方法で作成した.簡易防音室内で話者に日本語音節(Table2-1)を発話させ,音声と発話時の正面顔動画像をデジタルビデオに収録する.話者は日本語を母国語とする成人男女各 1 名ずつとした.音声はサンプリングレート 48kHz の 16bit モノラル

形式,顔動画像は標準NTSCカラーDV形式で記録する.次に収録した音声と動画像をデジタルビデオからWindows PCにキャプチャし,各音節区間をビデオ編集ソフトウェア（Adobe Premiere 6.0）により切り出して3〜4秒程度のオリジナル音節刺激を作成した（話者2名×101音節の合計202刺激）.

このオリジナル音節刺激から,9条件の実験刺激を作成した.まず刺激の音声に+∞, +10, 0, -10, -∞dBの5通りのS/N比で白色雑音を付加する.S/N=+∞はオリジナル刺激,-∞は音声を白色雑音に置換した刺激である.次にこれら5種類の刺激のうちS/N=-∞以外の4種類に対して,話者の顔動画像をブランク画像で置き換えた刺激を作成する.これらの刺激では音声情報だけが認識の手がかりとなる.刺激総数は101音節×9条件×2話者の1818個である.

	/a/	/i/	/u/	/e/	/o/	/ja/	/ju/	/jo/
-/-	あ	い	う	え	お			
/k/	か	き	く	け	こ	きゃ	きゅ	きょ
/s/	さ	し	す	せ	そ	しゃ	しゅ	しょ
/t/・/ts/	た	ち	つ	て	と	ちゃ	ちゅ	ちょ
/n/	な	に	ぬ	ね	の	にゃ	にゅ	にょ
/h/	は	ひ	ふ	へ	ほ	ひゃ	ひゅ	ひょ
/m/	ま	み	む	め	も	みゃ	みゅ	みょ
/y/	や		ゆ		よ			
/r/	ら	り	る	れ	ろ	りゃ	りゅ	りょ
/w/	わ				*を			
/g/	が	ぎ	ぐ	げ	ご	ぎゃ	ぎゅ	ぎょ
/z/・/dz/	ざ	じ	ず	ぜ	ぞ	じゃ	じゅ	じょ
/d/・/dz/	だ	*ぢ	*づ	で	ど	*ぢゃ	*ぢゅ	*ぢょ
/b/	ば	び	ぶ	べ	ぼ	びゃ	びゅ	びょ
/p/	ぱ	ぴ	ぷ	ぺ	ぽ	ぴゃ	ぴゅ	ぴょ
-								/n/ ん

*を記した音節,を≒お,ぢ≒じ,づ≒ず,ぢゃ≒じゃ,ぢゃ≒じゅ,ぢょ≒じょを同音とした.

Table2-1: 日本語基本音節

2.2 実験

被験者は簡易防音暗室内でヘッドホンを装着して椅子に座り,机上のdisplayを注視する.ヘッドホンから音刺激,displayからは話者動画像刺激が呈示される.被験者とdisplayの距離は90cm,画像サイズは17インチのCRTdisplay（解像度1024×768）に640×480（正方形ピクセル）である.この画像サイズは視野角で約10度に対応し,文字などの有効視野範囲として知られている凝視点から半径5度の傍中心窩に基づいている[5].実験中は机に固定した顔面固定器で被験者の顔の位置を固定し,視野角が変わらないようにした.音声は約80dB SPLの音量で呈示する.

各sessionは1条件,101音節の試行で構成される.2sessionごとに5-10分の休憩を設け被験者の注意レベルを保つ.session内での呈示順序はランダムである.これら9sessionで構成される実験を各被験者に対し5回行った.各被験者には合計9session×5回×101音節=4545試行の刺激が呈示される.また,各被験者に対して実験開始前日に2sessionずつのトレーニングを行っている.被験者は20歳台の男子大学生4名で,日本語を母国語とし全員が健康な聴覚を有する.

3.結果

3.1.音節認識率

各条件における認識率をFig.3-1に示す.A条件（音声のみ）では雑音レベルの増加に伴って音節認識率が低下する.雑音を付加しない（S/N=∞）場合の音節認識率はすべての被験者で100%近いが,S/N比が-10dBになると約37%まで低下する.これに対し視覚情報を追加したA+V条件では,A条件と同様雑音レベルの増加で認識率は低下するが,すべての雑音レベルでA条件より認識率が高かった.この認識率の差は,雑音レベルの増加に伴い大きくなり,S/N=-10dBにおいて約20%である.以上の結果から,雑音により認識が不完全になる場合に,話者の動画像のような視覚情報を併用することで音節認識率を改善することが可能であること,雑音レベルが大きいほどその効果が高くなることが分かる.

また視覚情報だけであるV条件の平均認識率は16.2%である.これは視覚情報だけでは音節認識に十分な情報を供給できないことを意味する.さらにS/N=-10dBにおけるA条件とA+V条件の認識率の差がV条件の認識率を上回ることから,A+V条件における音節認識率の改善は視聴覚情報の相補的統合により達成されている可能性が高いことが分かる.

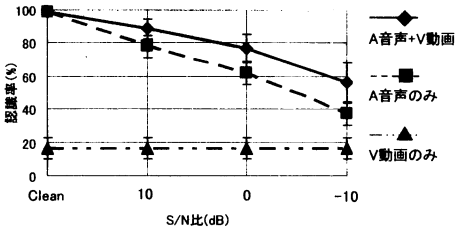

Fig3-1:音節認識率

3.2.母音の誤認パターン

Table3-1(A)〜(I)に母音の誤認地図を示す.表中の数字は刺激母音に対する回答母音の割合を百分率で示したものであり,全被験者回答の平均値である.この表から母音の認識率が非常に高いことが分かる.S/N=-10dBのとき母音の平均認識率はA条件で83.8%,A+V条件で97.5%であり,同じS/N比の音節認識率（A条件で37.3%,A+V条件で56.3%）と比べて非常に高い.またV条件でも音節の平均認識率16.2%に対して母音の平均認識率は81.9%で同様である.さらに全てのS/N比でA+V条件の平均母音認識率はA条件より高い.従って本実験において雑音レベルが増加した場合は,主に子音部分で認識の誤りが生じていると考えられる.

また表から母音の認識誤りのパターンがA条件とV条件で質的に異なることが分かる.A条件では/i/と/u/,/e/と/o/の間に誤りが生じやすく,/N/を/i/と誤認する場合も多い（Table3-1(A),(C),(E),(G)）..これに対してV条件では/a/と/e/,/u/と/o/の間で誤りが生じやすい（Table3-1(I)）.これは視覚と聴覚から得られる母音の手がかりが相補的であることを意味する.A+V条件の平均母音認識率がA条件やV条件より高くなるという結果は,この視聴覚の相補的な手がかりを適切に統合するメカニズムに起因するものと考えられる.

(A) A. Clean　ave. = 100.0

	a	i	u	e	o	N	*
a	100	0	0	0	0	0	0
i	0	100	0	0	0	0	0
u	0	0	100	0	0	0	0
e	0	0	0	100	0	0	0
o	0	0	0	0	100	0	0
N	0	0	0	0	0	100	0

(B) A+V, Clean　ave. = 100.0

	a	i	u	e	o	N	*
a	100	0	0	0	0	0	0
i	0	100	0	0	0	0	0
u	0	0	100	0	0	0	0
e	0	0	0	100	0	0	0
o	0	0	0	0	100	0	0
N	0	0	0	0	0	100	0

(C) A, S/N=+10dB　ave. = 96.5

	a	i	u	e	o	N	*
a	100	0	0	0	0	0	0
i	0	94	3	0	0	1	2
u	0	0	100	0	0	0	0
e	0	0	0	100	0	0	0
o	0	15	0	0	85	0	0
N	0	0	0	0	0	100	0

(D) A+V, S/N=+10dB　ave. = 99.8

	a	i	u	e	o	N	*
a	100	0	0	0	0	0	0
i	0	100	0	0	0	0	0
u	0	0	100	0	0	0	0
e	0	0	0	100	0	0	0
o	0	0	0	0	100	0	0
N	0	0	0	0	0	100	0

(E) A, S/N=0dB　ave. = 96.7

	a	i	u	e	o	N	*
a	100	0	0	0	0	0	0
i	0	92	3	0	0	1	4
u	0	0	96	0	0	0	3
e	0	0	1	97	0	0	1
o	0	0	0	0	100	0	0
N	0	5	0	0	0	95	0

(F) A+V, S/N=0dB　ave. = 99.3

	a	i	u	e	o	N	*
a	99	0	0	0	0	0	1
i	0	100	0	0	0	0	0
u	0	0	100	0	0	0	0
e	0	0	0	98	0	0	2
o	0	0	0	0	100	0	0
N	0	0	0	0	0	100	0

(G) A, S/N=-10dB　ave. = 83.8

	a	i	u	e	o	N	*
a	98	1	0	0	0	0	1
i	0	83	10	0	0	1	7
u	0	12	84	0	0	0	4
e	0	1	0	89	5	0	5
o	1	0	0	16	81	0	3
N	0	30	0	0	0	70	0

(H) A+V, S/N=-10dB　ave. = 97.5

	a	i	u	e	o	N	*
a	98	1	0	0	0	0	1
i	0	98	0	0	0	0	3
u	0	0	97	0	0	0	2
e	0	0	0	96	0	0	3
o	0	0	0	0	97	0	3
N	0	0	0	0	0	100	0

(I) V　ave.= 81.9

	a	i	u	e	o	N	*
a	83	0	0	8	0	0	8
i	0	92	0	2	0	0	5
u	0	0	81	0	9	0	10
e	43	0	0	45	0	0	11
o	0	0	1	0	90	0	8
N	0	0	0	0	0	100	0

Table3-1: 母音認識の confusion matrix
1 列目は呈示刺激母音,1 行目は回答母音

3.3.子音の誤認パターン

Fig.3-2 に子音間の誤認を模式図にしたものを示す.図では 10%以上の誤認があった子音間の誤認率を刺激子音から回答子音への矢印で表現している.両方向の矢印は,二つの子音間での相互誤認を意味する.数値は誤認率,矢印の太さは誤認率の高さに対応する.

S/N=-10dB における平均子音認識率は A 条件で 46.2%,A+V 条件で 62.5%であり,Fig3-1 と比較参照すれば雑音レベルの増加に伴う音節認識率の低下が主に子音の誤認に起因するということがわかる.

また同じ条件での誤認パターンは,雑音レベルに依らずほぼ一定であることが分かる.S/N=+10dB や 0dB の誤認パターンは基本的に S/N=-10dB におけるパターンの矢印の太さを細くしたものである.この傾向に当てはまらないものは,A 条件の S/N 比=-10dB から S/N 比=0dB の/w/→/g/のみである.この結果は,子音の誤認が無作為に生じるものではなく,雑音環境下で誤りが生じ易いいくつかの子音群が存在することを意味する.

また Fig.3-2 で同じ S/N の A 条件と A+V 条件を比較すると,視覚情報による正の効果は/m/⇔/n/,/p/⇔/k/,/p/⇔/h/,/r/⇔/b/,/b/⇔/g/の誤認が消滅することである.これらの誤認は A 条件で雑音レベルを下げてもさほど減少しないが,同じ S/N でも A+V 条件にすることで劇的に減少する.一方,視覚情報による負の効果は,A+V 条件の S/N=-10 dB において/m/⇔/p/, /y/⇔/g/, /t/⇒/g/の誤認が表れることである（Fig.3-2 の*印）.これらの誤認は同じ S/N の A 条件では殆ど見られず,V 条件でも多くはない（/m/→/p/が 19%で例外）.よって A 条件でも V 条件でも生じない誤認が A+V 条件で生じるということになる.この結果は,被験者の視聴覚情報の統合過程が単純な線形系では説明できないことを意味する.

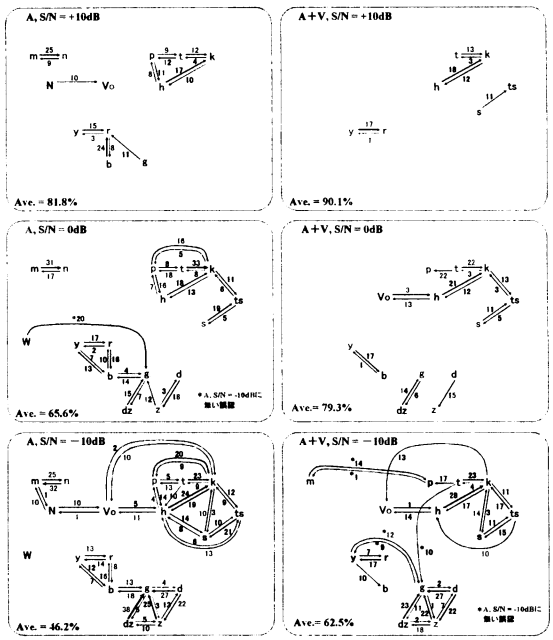

Fig.3-2: 子音間の誤認模式図

4.考察

4.1.視覚情報による母音誤認

母音の誤認パターンが A 条件と V 条件で質的に異なることの原因について考察する.Fig.4-1 に調音論的に抽象化された母音の関係を示す.水平方向が舌の位置（調音位置）に,垂直方向が顎の開き（母音の高さ）に対応する.実線で囲まれた母音群は V 条件で,点線で囲まれた母音群は A 条件で混同の多いものに対応する.

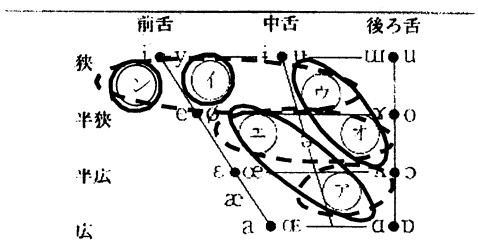

Fig.4-1:聴覚情報と視覚情報の母音弁別

図から視覚情報で弁別可能な母音群と聴覚情報で弁別可能な母音群が相補的な関係であることが分かる

図から視覚情報によって弁別可能な母音の 4 つのグループ(/a/,/e/), (/u/,/o/), (/i/), (N)を弁別する手がかりは水平方向に対応する調音位置であるように見える.し

かし調音論における調音位置とは舌の口腔内での前後位置に対応するため,視覚でこの様な情報が検出されているとは考え難い.逆に視覚で検出し易いはずの顎の開きでは,上記弁別グループを上手く説明できない.この問題は,"顎の開閉"と"口の開閉"は単純に対応していないと考えると解決する.Fig.4-2 に定常母音発話時の X 線写真から得られた声道断面の形状を示す[7].この図から鉛直方向の口の開きは,/u/→/o/→/i/→/e/→/a/の順で大きくなる.また音節/N/では口は開かない.この順で隣り合う/u/と/o/,及び/a/と/e/が混同されたと考えると視覚情報による4つのグループを説明できる.

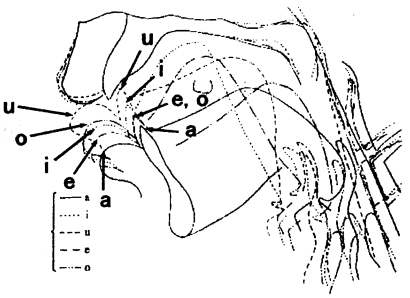

Fig.4-2: 母音発話時の声道形状

この仮説を検証するために,本実験で用いた2名の話者の5母音発話画像から鉛直方向の口の開きを測定した(Fig.4-3).結果は前述の仮説とほぼ整合するが,いくつか問題もある.第一は,男性話者の/i/の口の開きが/e/より大きいため V 条件のには/i/-/a/間の混同が発生するはずだが,実際にはそのような結果は得られていない.第二は,両方の話者で/a/-/e/間や/o/-/u/間の口の開きの差が,/e/-/i/や/i/-/o/の差と大きく変わらないため口の開きの順番が隣り合う2つの母音間全てで混同が生じる可能性があり,(/a/,/e/), (/i/), (/u/,/o/)のグルーピングが発現するメカニズムを説明できない.さらに第三は,本実験では/e/から/a/への混同が顕著だが,その逆は殆どないことが挙げられる.被験者が視覚情報から得られる手がかりには,鉛直方向の口の開き以外に口の幅,唇の形状,歯の見え方(顎の開き方),唇周囲筋の動きなどが考えられる.被験者はこれらの動的な変化や複数の視覚的手がかりを複合的に分析して読唇すると思われる.これらの手がかりのうち何がどの程度音声認識に影響するかを調べることは今後の課題である.

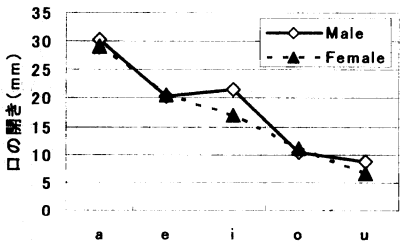

Fig.4-3: 口の開き

4.2.聴覚情報による母音誤認

Fig.4-1 から雑音環境下の A 条件による誤認は主に調音位置に対して生じていることが分かる.即ち,聴覚情報だけでは前舌・後舌母音の弁別(調音位置)は出来ないが,狭母音・広母音の弁別(顎の開き)は可能である.一般に顎の開きは第 1 ホルマント周波数に,調音位置は第2ホルマント周波数に対応すると言われている.よって今回の実験では雑音により第2ホルマント周波数がマスクされたのではないかと予測される.

Fig.4-4 に母音刺激のスペクトルを示す.図(A)-(E)は S/N=∞(雑音なし),(F)-(J)は S/N=-10dB の刺激の母音である.図(A)-(E)では,ホルマントピークが明確に見られる.図(F)-(J)の雑音を付加した刺激では,このようなスペクトル形状が殆ど失われている.僅かに確認できるのは第1ホルマントピークだけで,第2ホルマントピークは埋もれて確認することが出来ない.雑音により第2ホルマントピークがマスクされる事により聴覚情報だけで前舌・後舌母音の弁別が出来ないと考えることができる.この仮説が正しければ,本実験で用いた白色雑音ではなく高周波数ほどエネルギーが小さいピンクノイズを用いた場合に同じ S/N で前舌・後舌母音の弁別能力が改善すると考えられる.この仮説の信頼性については,今後の実験で明らかにしていきたい.

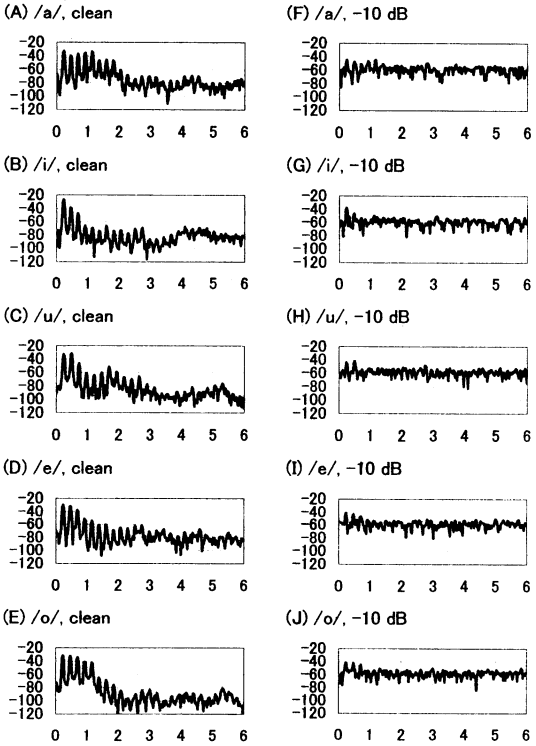

Fig.4-4:母音刺激のスペクトル

4.3.視覚情報による子音誤認

V 条件の実験結果から視覚情報で弁別可能な子音は,①[/N/],②[/p/,/b/,/m/,/w/],③[その他の子音]の3グルー

プに分けられる.このグループ分けは発話時の唇の動きから容易に予測できるもので先行研究とも良く一致する.①のグループでは発話の最初から最後まで唇は閉じたままである.このような音節は/N/以外存在せず,V条件での認識率は100%になっている.②のグループは,最初に唇がいったん閉じてから開く.V条件の実験ではこのグループに属する4つの子音間の誤認は多いが,異なるグループの子音への誤認は殆ど生じない.③のグループは,唇が閉じない状態から発話が開始される.このグループに属する子音が最も多い.

3.3で述べた様に視覚情報統合による正の効果は/m/⇔/n/,/p/⇔/k/,/p/⇔/h/,/r/⇔/b/,/b/⇔/g/の誤認が減少することだが,これらは全て上記②と③のグループに属する子音の誤認である.A条件で多いこれらの子音間の誤りがA+V条件で減少する原因は,発話開始時の話者の唇の開閉に関する情報を得るためと考えられる.

一方,視覚情報による負の効果は,S/N = -10 dBのA+V条件で/m/⇔/p/, /t/⇒/g/,/y/⇔/g/の誤認の増加として現れる.そこで画像付加により,音素/m//mj/に該当する160回分の刺激に対する誤認パターンがどのように変化するかを調べる（Fig.4-5a）.グラフはS/N=-10dBのデータで,縦軸の数値はA+V条件とA条件の誤認率の差分である.この図から/m/⇒/n/の誤認は改善されていることが分かる.また/m/⇒/p/の誤認が生じるのは主に女性話者の刺激である.この誤認パターンは話者依存の強いものであるが/m/と/p/はいずれも上記②のグループの“唇音”であり口の動きが似ている.

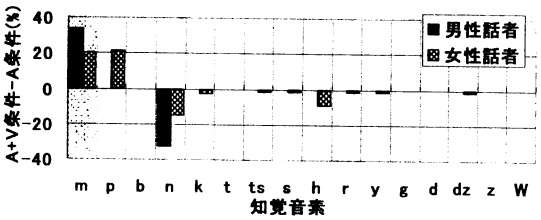

Fig.4-5a:画像付加による知覚変化
[/m//mj/呈示,S/N=-10dB]

次に/t/⇒/g/について考える.これらは③のグループに属する子音である.子音/t/を含む刺激音節に対する誤認変化をFig.4-5bに示す.刺激全60回のうち誤認されたのは43回（約72%）で非常に多い.このうち誤認率10%を超えるのは,/t/⇒/k//kj/（14回）, /t/⇒/p//pj/（10回）, /t/⇒/g//gj/（6回）である.いずれも③のグループに属する/t/⇒/g//gj/の誤認は,A条件で/ta/⇒/ga/（2回）, /te/⇒/ge/（1回）だったものが,A+V条件では/ta/⇒/ga/（3回）, /to/⇒/go/（3回）と増加している.これらも唇の動きが似ている子音である.子音/k/への誤認は視覚情報の統合により男性話者で増大し,女性話者で減少している.また唇の動きが異なる/p/,/pj/への誤認は男性話者で減少し女性話者で増大している.男性話者の結果では唇の動きが異なる子音同士の誤認が改善され,似ている子音の誤認が増加した.女性はこれと反対の結果となった.なぜこのような結果になるのかはいまのところ不明である.

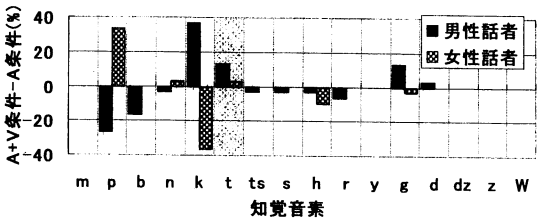

Fig.4-5b:画像付加による知覚変化
[/t/呈示,S/N=-10dB]

最後に/y/⇔/g/の誤認について考える.これらはともに③のグループに属する子音である.子音/y/を含む刺激音節に対する誤認をFig.4-5cに,また子音/g//gj/を含む刺激音節に対する誤認をFig.4-5dに示す./y/を含む刺激音節全60回のうち,誤認が認められたのは24回である.この24回のうち女性話者音声の誤認が22回と90%以上を占める.ここでも/m//mj/の場合と同様に話者依存性が見られる.また24回の誤認のうち誤認率10%を超える/y/⇒/g//gj/,/b//bj/の誤認は13回で54%を占める.これらは/y/⇒/gj/（7回）,/y/⇒/bj/（6回）ですべて日本語の「や,ゆ,よ」にあたる開拗音/j/の音素を含んだ音節に誤認し,母音はすべて一致している.ちなみに/bj/は上の②のグループであり,視覚情報を用いないA条件と比べると誤認は減少している.一方/g//gj/の音素を含む刺激音節全160回のうち,誤認が認められたのは65回である.このうち誤認率10%を超えるのは/g//gj/⇒/dz//dzj/の17回のみで,次いで/g//gj/⇒/y/（14回）,/g//gj/⇒/b//bj/（13回）となる.このうち/bj/については,上と同様A条件と比べて誤認は減少している.

以上のように,視覚情報の統合による負の効果は話者依存性が高い.しかしながら,唇の動きが似ている子音間で誤認が起き易くなっている現象も観測できた.また正の効果は上記の子音3グループの分類により良く説明することができる.

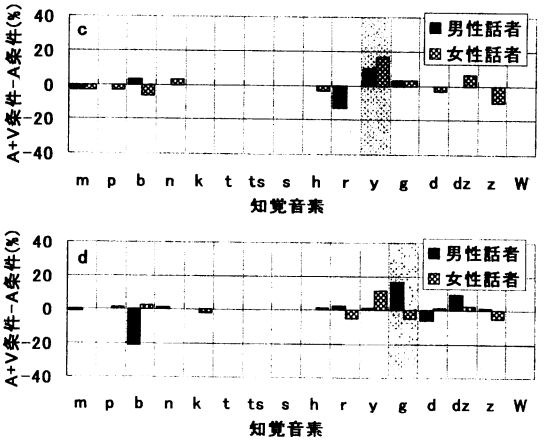

Fig.4-6c,d:画像付加による知覚変化
[c:/y/呈示,d:/g//gj/呈示,S/N=-10dB]

4.4.聴覚情報による子音誤認

S/N=-10dB の A 条件の結果から,聴覚情報で弁別可能な子音は，Ⅰ [/N/,/m/,/n/,/p/,/h/,/k/,/t/,/s/,/ts/], Ⅱ [/b/,/r/,/d/,/y/,/g/,/dz/,/z/], Ⅲ[/w/]の 3 グループになる.Ⅰは無声子音,Ⅱは主に有声子音で構成されている.ただしⅠには有声子音である鼻音（/N/,/m/./n/）も属す.

このような分類となる原因で考えられるのが,音節を構成する子音部分と母音部分のエネルギーの差である.一般に無声子音はエネルギーが小さく,有声子音は大きい.白色雑音を付加することで無声子音部分は雑音に埋もれてしまうが,有声子音部分は検出可能である.また鼻音は音韻論的に有声子音に分類できるがエネルギーは小さいことが知られている.従って被験者が子音部分のエネルギーの有無により子音を認識していると考えると,少なくともⅠとⅡのグループができる原因は理解できる.この仮説を検証するため,刺激の母音区間と子音区間の平均RMSの比を計算した.Fig.4-6は,上の子音グループごとにRMS比の平均値を,男性話者,女性話者,男女混合のそれぞれについて計算したものである.この図からⅠの子音区間に対する母音区間のRMS比は,話者に関わらずⅡよりも大きくなっていることが分かる.これは上の仮説を支持する結果である.しかしⅢに属する/w/がⅠとⅡの中間程度の RMS 比を持つことを考えると上の仮説はあくまで一次近似に過ぎない.また 4.2 で論じたように雑音の種類によって子音のグルーピングが変化する可能性も否定できない.雑音下で聴覚情報により生じる誤認パターンは,今後詳細な実験を行って明らかにして行く必要がある.

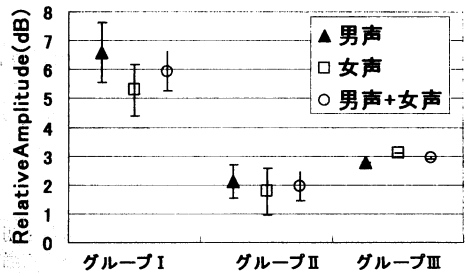

Fig.4-6:母音区間と子音区間の RMS 比率

4.5.子音誤認パターン

前節までの考察を踏まえて,視覚情報,聴覚情報に基づく子音の誤認パターンを模式化すると Fig.4-7 のようになる.実線が[A 条件, S/N = -10 dB]で 10%以上の誤認があったもの,同様に点線が[V 条件]で 10%以上の誤認があったものを結んだものである.横軸の視覚情報は,4.3 で述べた発話時の口の動きに対応している./N/は唇閉,/m//p//W//b/の両唇音に分類される 4 つの音素を含む子音は唇閉⇒唇開,それ以外の子音は唇半開⇒開という動きを示す.縦軸の聴覚情報は,4.4 で述べた聴覚情報から弁別可能である 3 つの子音グループに対応している.これらの情報が相補的であるために,A+V 条件ではA条件やV条件よりも子音の認識率が向上すると考えられる.また視覚情報による子音の分類は一般的なものであると思われるが,聴覚情報による分類は雑音の特性等により大きく変化する可能性がある.

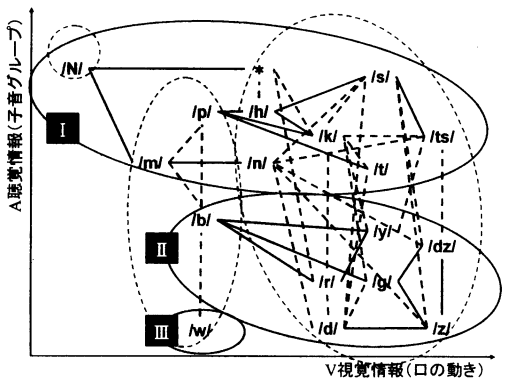

Fig.4-7:子音の誤認マップ

5.結論

(1)白色雑音を付加した音節を聴覚情報だけから認識する場合,S/N 比が減少すると認識率は単調に低下する.雑音環境下での認識誤りの多くは音節の子音部分で生じていると考えられる.(2)白色雑音を付加した音節を視聴覚情報から認識する場合,認識率は聴覚情報だけより高くなる.視覚情報の効果は,特に S/N が低い場合に大きくなる傾向がある.(3)音節を視覚情報だけから認識する場合,その認識率は非常に低い.(4) 母音の混同は,視覚情報から得られるパターンと高雑音（S/N=-10dB）の聴覚情報から得られるパターンが相補的になっている.視覚情報では唇の鉛直方向に開く大きさが,聴覚情報では第 1 ホルマント周波数が母音認識の手がかりとなっていると考えられる.(5) 子音の混同も,視覚情報から得られるパターンと,高雑音（S/N=-10dB）の聴覚情報から得られるパターンが相補的になっている.視覚情報では唇の動き方が子音認識の手がかりとなっていると考えると混同パターンがほぼ説明できる.

文　献

[1] 奥村晃弘, 宮崎敏彦 (1997) “唇の動き情報による騒音環境下での音声認識性能の改善” 情報処理学会研究報告 V01.97, N0.88(HI-74), pp. 43-48

[2] C.Neti et al. (2000) "Audio-visual speech recognition"CLSP Workshop 2000, final report

[3] 積山薫 (1997) “顔と声による音声知覚” 信学技法 PRMU97-140, HIP97-21, pp.83-90

[4] J.Luettin (1997) “Visual speech and speaker recognition” Univ. of Sheffield dissertation

[5] 大山正, 今井省吾, 和気典二 編“新編 感覚・知覚ハンドブック” 誠信書房, pp918-930, 1996

[6] 中田和男 (1995)“音声” コロナ社

[7] 大友信一 (1979)“X 線映画資料による母音の発音の研究” 国立国語研究所報告 60